Real-time Reproduction of 3D Human Images in Virtual Space Teleconferencing

Jun OHYA, Yasuichi KITAMURA, Haruo TAKEMURA,

Fumio KISHINO and Nobuyoshi TERASHIMA

ATR Communication Systems Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02, Japan
Email: ohya@atr-sw.atr.co.jp

Abstract

For the first time, real-time reproduction of a 3D human image is realized by the experimental system the authors recently built for the realization of a virtual space teleconferencing, in which participants at different sites can feel as if they are at one site and can cooperatively work. In the teleconferencing system, the 3D model of a participant is constructed by a wire frame model mapped by color texture and is displayed on the 3D screen at the receiving site. In the experimental system, to realize real-time detection of facial features at the sending site, tape marks are attached to facial muscles, and the marks are tracked visually. To detect movements of the head, body, hands and fingers in real-time, magnetic sensors and data glove are used. When the movements of the participant are reproduced at the receiving site, the detected results are used to drive the nodes in the wire frame model. Using the experimental system, the optimum number of nodes for real-time reproduction is obtained. Promising results for real-time cooperative work using the experimental system are demonstrated.

1 Introduction

The authors are aiming at realizing **VI**rtual **S**pace **TEL**econferencing(VISTEL)[1], in which the teleconference participants at different sites can feel as if they are all at one site, allowing them to hold meetings and work cooperatively. In VISTEL, the 3D models of the participants at different sites are combined into an artificially created 3D image of a virtual space, and by displaying the 3D image of the virtual space on the 3D screen at each site, the participants will get the feeling of meeting each other in a common space. The 3D modeling in VISTEL makes it possible to realize motion parallax essential to natural 3D display and allows participants to achieve eye contact that would occur in actual conferences. To achieve naturally smooth communication among the participants, it is necessary to reproduce their movements at the 3D models in real-time.

At each site in VISTEL, the 3D models of the participants at the other sites have been created by mapping the color texture of the participants' bodies onto the 3D wire frame models and are then displayed on the 3D screen. Therefore, to reproduce the facial expressions and body movements in real-time at the receiving site, it is necessary to detect these motions of the participants at the sending site in real-time. There have been research works on facial expression synthesis by transforming the 3D wire frame models for human faces[2],

0-7803-1363-1/93 $3.00 © 1993 IEEE

[3], but those projects have not yet realized real-time facial expression reproduction. Moreover, no work has yet dealt with reproducing movements of the upper half of the body as well as that of the head.

In this paper, an experimental system is constructed to reproduce the movement of the upper half of the human body in real-time. The following sections describe more details of the experimental system. Then, by using the experimental system, the conditions that achieve real-time natural and high quality reproduction are investigated. Finally, promising results for real-time cooperative work using the experimental system are demonstrated.

2 Real-time Reproduction of 3D Human Images

2.1 Outline

The experimental system for VISTEL is constructed so that two persons can participate. Figure 1 shows a teleconference scene, where persons A and B are at Sites #1 and #2, respectively. At Site #1, for example, the 3D image of person B is placed in an artificially created virtual space (cooperative work space). The virtual space, including the image of person B, is displayed on the 3D screen at Site #1 so that person A can cooperatively work with person B.

To realize natural interaction between persons A and B, the movement of person B needs to be detected and displayed in real-time. To achieve this, as shown in Fig.2, the experimental system consists of three modules: 3D modeling of the human body, real-time detection of the human body's motion and real-time synthesis of the human image. In the following subsections, each module is explained.

2.2 3D modeling of the human body

As described above, the 3D images of participants are modeled in advance. This 3D modeling is done for each part of the human body.

As shown in Fig.3, we use the Cyberware Color 3D digitizer, which rotates around an object, projects laser stripes, and acquires color information and 3D coordinates of each point on the surfaces of the parts. By using a utility software, the acquired 3D coordinate data are transformed to a 3D wire frame model, which consists of triangular patches. More specifically, if the difference in the directions of the surface normals to neighboring points is smaller than the threshold, the points are merged to form an area with a uniform direction. After this process is applied to all of the points, each area is divided into triangular patches. Then, the color information of each point is mapped to its corresponding triangular patch.

As shown in Fig.3, the 3D model of each part is articulated with another.

2.3 Real-time detection of human body motion

Image processing could be useful for motion detection in this system because of its passive nature. However, since an existing computer must be used, it is very difficult to realize real-time detection of body motion with most image processing technologies. Therefore, as shown in Fig.2, the following alternatives are used in the current implementation.

For real-time detection of facial expressions, nine blue tape marks are attached to the facial muscles that can strongly affect changes in facial expressions. The marks are then tracked in the images acquired by a TV camera positioned in front of the face. The images are thresholded to discriminate the blue marks, and for each mark a window is set to keep the search area small and to avoid errors in tracking. The position of a mark is obtained

from the center of gravity of the blue pixels in the window. In the current implementation, for stable mark tracking, person B at Site #2 in Fig.1 wears the helmet shown in Fig.2. The TV camera is mounted to the rim fixed to the helmet so that images of faces are obtained from the same point of view.

On the rim of the helmet, a bulb is also fixed so that the TV camera can detect the virtual images of the bulb created by reflectances from the corneas. Tracking is performed within the windows set at the eyes. With the tracking results, gaze and blink detection can be done.

To detect the positions of the head, hands and body in real-time, magnetic sensors[4], which can detect the three translation and rotation parameters, are used. The magnetic sensors are attached to the helmet, the backs of the hands and the area below the throat. To detect finger movements in real-time, the participant wears a data glove[5], which can detect bends of the fingers.

2.4 Real-time synthesis of human images

The results of tracking the marks on the face are used to move the nodes of the wire frame model of the face. Figure 4 shows the marks on the face and their corresponding nodes in the model. Although a face is inherently a 3D object, the movement of the marks are detected as 2D motion in the images from the TV camera. Therefore, knowledge and constraints are necessary for driving the 3D model.

In this paper, the mark on the nose is assumed to be the immovable reference point, and the distances between each mark and the reference point are measured in advance. If changes in the distances arise according to changes in facial expressions, the 3D model of the face is driven, based on the detected movement of the marks. An example of the transformations of the face model is as follows:

Based on the movement of the mark on the lower lip, the nodes belonging to the lower jaw are moved. When the movement of the mark is small, the mouth is judged to be narrowly opened, and then the nodes in the lower jaw, including the lips, are driven on the cylinder whose central axis is the line connecting the joints between the upper and lower jaws on both sides of the face. When the mark movement is large, the mouth is judged to be wide-open, and the lower jaw moves down in the vertical direction.

The results of tracking the virtual images of the bulb are used to determine the gaze of the synthesized eyes. If the virtual image is not detected in the window, the eye is judged to close.

The 3D coordinates of the nodes in the parts for the head, hands, fingers and body are obtained from the information detected by the magnetic sensors and data glove. The 3D coordinates of the nodes involved in each articulation are calculated by a linear interpolation based on the coordiantes of the parts on both sides of the articulation.

3 Configuration of the Experimental System

As shown in Fig.5, the WS(Iris Crimson, Reality Engine) at Site #1, which is characterized by its fast texture mapping function, is used for real-time 3D display of person B at Site #2. The 3D human image in a synthesized virtual space is output to a 70-inch stereoscopic display[6], and person A at Site #1 can work cooperatively with person B by wearing LCD shutter glasses for stereo viewing[6]. The authors are now studying a 3D display that does not require special glasses[7].

Facial expressions of person B are detected by a WS(Iris 4D340/VGX) at Site #2. Images acquired by the color TV camera attached to person B's helmet shown in Fig.2 are input to the two thresholding devices: a chroma keyer for the blue tape marks and a level keyer for the virtual images of the bulb. Then, the tracking process is carried out by the WS.

Another WS(Iris 4D240/VGX) at Site #2 is used to process the data from the data glove and magnetic sensors of person B. At Site #2, the same virtual space as Site #1 is created by the WS. However, due to its limited texture mapping ability, the human image of person A is replaced by a simple CG image.

Person B's motion information acquired by the two WS's is sent to the WS at Site #1 through the Ethernet and drives the 3D model of person B.

Microphones and speakers are at both sites so that person A and B can communicate with their voices. To synchronize voices with image, there is delay equipment on the acoustic line between the two sites.

4 Experimental Results and Discussion

Experiments on real-time detection of facial expressions were carried out. Tracking the blue tape marks and virtual images of the bulb for various facial expressions allowed tracking at a speed of 11 frames/sec. The data glove and magnetic sensors can achieve faster detection than this.

For high quality human image display, the number of nodes in a wire frame model should be large, but the large number of nodes could make real-time reproduction impossible. Therefore, it is necessary to clarify the optimum number of nodes. As described in **2.2**, the number of nodes in a wire frame model can be changed according to the threshold value for the merging process. Human head models with 203, 506, 995, 1985, 5005 and 9556 nodes and a model for the upper half of the body with 5300 nodes were created. The display speed was measured for each model, where time necessary for movement detection is involved. The results are indicated in Fig.6. When only the head is displayed, the speed is quite fast. However, with the body and background(objects in the virtual space), the speed is much slower. This is caused by access to the converters for the data glove and magnetic sensors through Ethernet.

By observing each reproduction, the image quality gets better with an increase in the nodes, but it is not improved if the node number exceeds 1000. With 1000 nodes for the head, 6 frames/sec displaying speed can be achieved, as shown in Fig.6, even when the virtual space is also displayed. We conducted VISTEL using the 1000 node head model(Fig.7). Consequently, participants' evaluations agreed that the 3D human image reproduction is quite natural and smooth, but further study is necessary for more accurate image quality evaluation.

5 Conclusions

Real-time reproduction of human 3D images in VISTEL has been studied. The 3D model of a participant was created by a wire frame model mapped by color texture. The detected movement information drives the 3D model. An experimental system was constructed to detect and reproduce the movement of a participant in real-time. By using the experimental system, it was possible to obtain the optimum number of nodes in a wire frame model. Accordingly, the model with 1000 nodes for the head and 5300 nodes for the upper half of the body can be reproduced at 6 frames/sec, which can be evaluated as natural and smooth reproduction.

In the current implementation, to achieve real-time reproduction, tape marks, a data glove and magnetic sensors are used. However, these tools are not appropriate for natural human communications. It is our goal to realize a system that does not require such tools.

References

[1] F.Kishino "Communication with realistic sensations through integrated multi-media environment", Proc. of Language and Vision Workshop(1st US-Japan Workshop on Integrated Systems in Multi-media Environments), pp49-58, (Dec. 1991)

[2] S.Morishima et al., "Model-based facial image coding controlled by the speech parameter", PCS88, 4.4 (1988)

[3] D.Terzopoulos et al., "Physically-based facial modelling, analysis and animation", The Journal of Visualization and Computer Animation, Vol.1, pp.73-80 (1990)

[4] F.H.Raab et al., "Magnetic position and orientation tracking system", IEEE Tr. on AES-15, 5, pp.709-718 (1979)

[5] T.G.Zimmerman et al., "Hand gesture interface device", in CHI+GI'87, pp.189-192 (1987)

[6] H.Takemura et al., "Cooperative work environment using virtual workspace", Proc. of CSCW'92, pp.226-232 (1992)

[7] N.Tetsutani et al., "Stereoscopic display method employing eye-position tracking and HDTV LCD-projector", International Workshop on HDTV'92, Vol.II, pp.60-1 - 60-8 (1992)

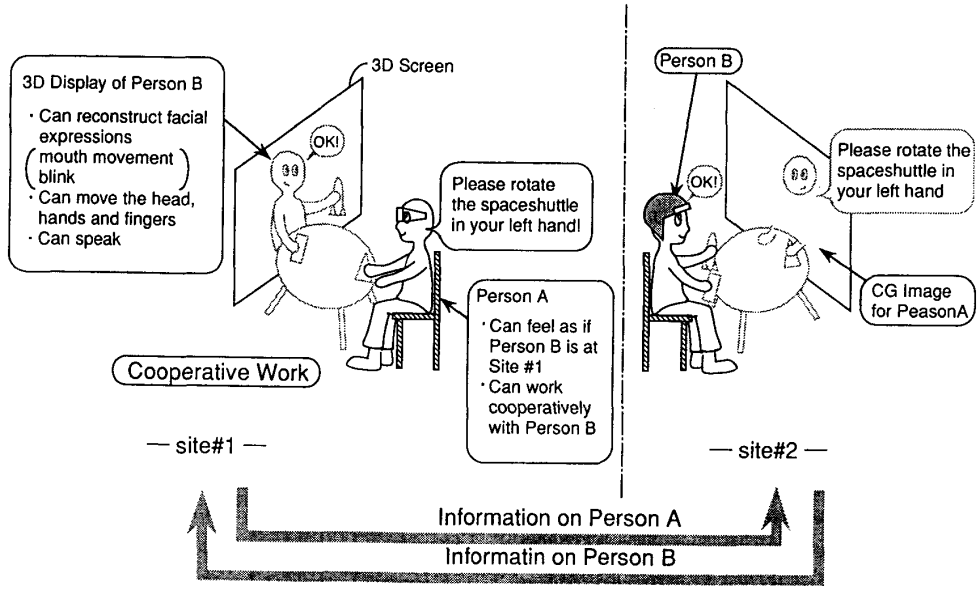

Fig.1 Virtual Space Teleconferencing

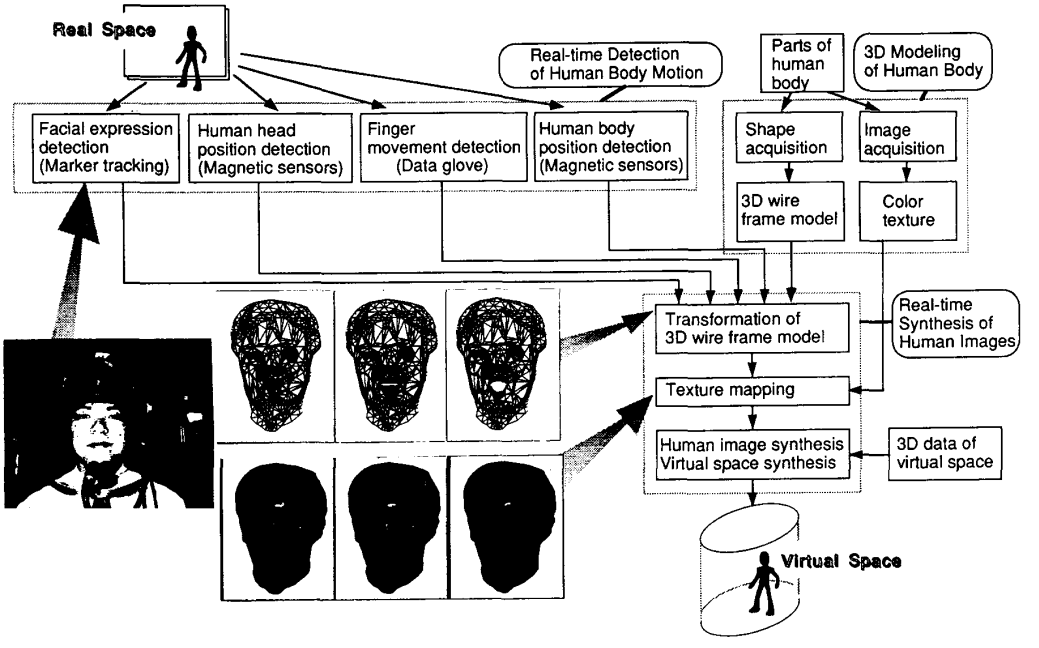

Fig.2 Real-time Detection and Synthesis of 3D Human Images

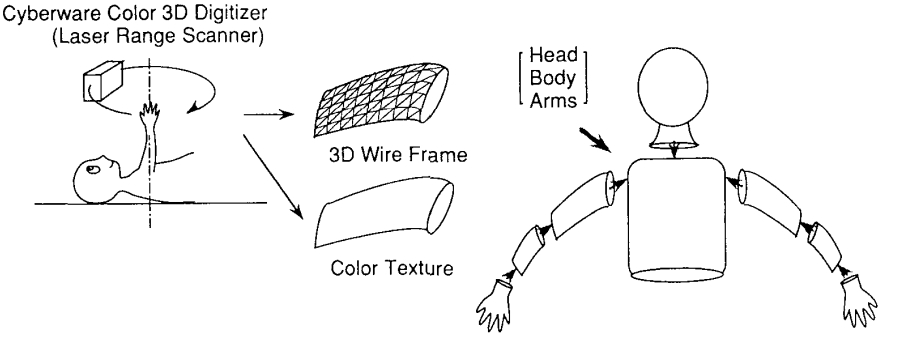

Fig.3 3D Modeling of Human Body

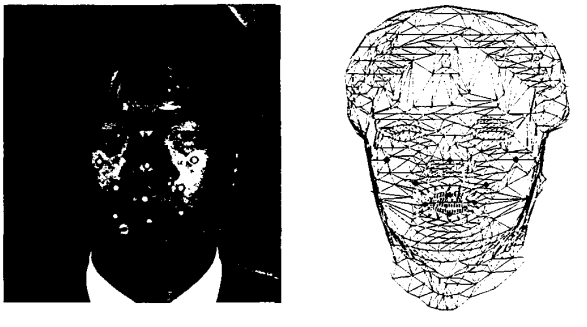

Fig.4 Mark Positions and Their Corresponding Nodes

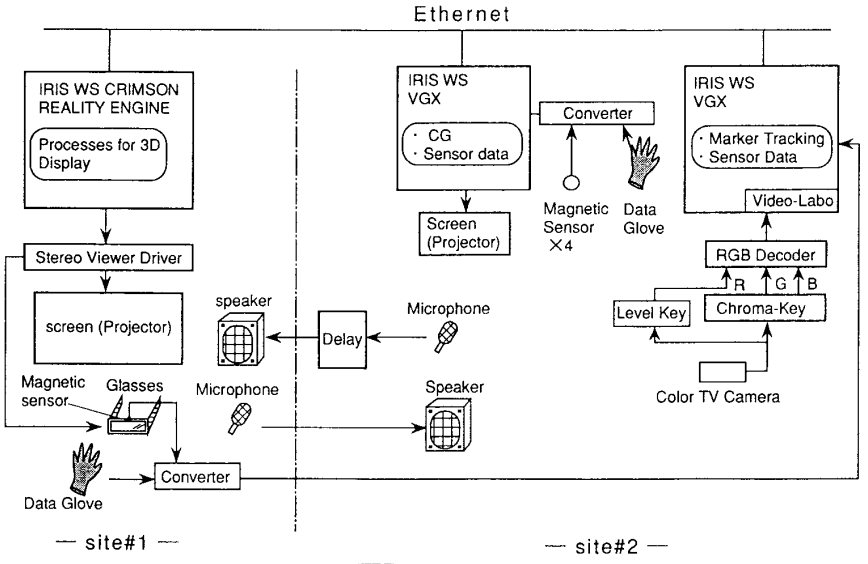

Fig.5 Configuration of the Experimental System

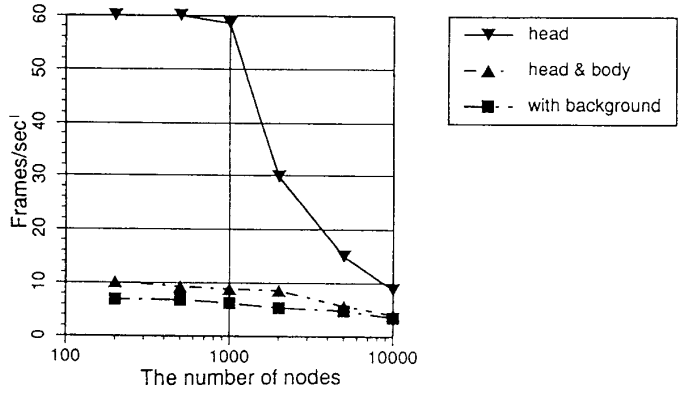

Fig.6 Relationsip between the Number of Nodes and Display Speed

Fig.7 Scene of Virtual Space Teleconferencing